U-NEED：以用户需求为中心的电子商务对话式推荐细粒度数据集

Yuanxing Liu
leeeeoliu@gmail.com
独立研究者
中国黑龙江哈尔滨

Weinan Zhang*
weinanzhang@gmail.com
独立研究者
中国黑龙江哈尔滨

董宝华
baohua.dbh@alibaba-inc.com
阿里巴巴集团
中国浙江省杭州市

严帆
fanyan.fy@alibaba-inc.com
阿里巴巴集团
中国浙江省杭州市

杭旺
hwang.rvec@gmail.com
独立研究者
中国黑龙江哈尔滨

范峰
fengfan.fengfan@alibaba-inc.com
阿里巴巴集团
中国浙江杭州

Yifan Chen
chenyifan.stu@gmail.com
独立研究者
中国黑龙江哈尔滨

Ziyu Zhuang
royzhuang1124@gmail.com
独立研究者
中国黑龙江哈尔滨

Hengbin Cui
alexcui.chb@alibaba-inc.com
阿里巴巴集团
中国浙江杭州

Yongbin Li
shuide.lyb@alibaba-inc.com
阿里巴巴集团
中国浙江杭州

Wanxiang Che
wanxiang@gmail.com
独立研究者
中国黑龙江省哈尔滨市

摘要

对话式推荐系统（CRSs）旨在理解对话中表达的信息需求和偏好，并向用户推荐合适的商品。现有的大多数对话式推荐数据集都是合成的或通过众包模拟生成的，与真实场景存在较大差距。为弥合这一差距，先前的工作基于电子商务场景中用户与客服人员之间的售前对话，构建了数据集 E-ConvRec。然而，E-ConvRec 仅提供粗粒度标注，以及用于在售前对话中进行推荐的通用任务。与之不同，我们以真实用户需求为线索，探索复杂售前对话中的电子商务对话式推荐，即以用户需求为中心的电子商务对话式推荐（UNECR）。

本文构建了一个来自真实电子商务场景、以用户需求为中心的电子商务对话推荐数据集 （U-NEED）。U-NEED 包含 3 类资源：（i）7,698 fi个经过细粒度标注的售前对话，涵盖 5 个主要类别；（ii）333,879 条用户行为；以及（iii）332,148 个产品知识三元组。为推动 UNECR 研究，我们提出了 5 项关键任务：（i）售前对话理解；（ii）用户需求引导；（iii）基于用户需求的推荐；（iv）售前对话生成；以及（v）售前对话评估。我们为每项任务建立了基线方法和评估指标，并报告了 U-NEED 上 5 项任务的实验结果。此外，我们还报告了 3 个典型类别上的结果。实验结果表明，UNECR 在不同类别中面临的挑战各不相同。

CCS 概念

• **计算方法**！**话语、对话与语用**；• **信息系统**！**网络日志分析**。

关键词

对话式推荐、用户需求、对话语料库

ACM参考文献格式：

Yuanxing Liu、Weinan Zhang、Baohua Dong、Yan Fan、HangWang、FanFeng、Yifan Chen、Ziyu Zhuang、Hengbin Cui、Yongbin Li和Wanxiang Che。2023。U-NEED：以用户需求为中心的电子商务对话式推荐细粒度数据集。载于第*46*届国际*ACM*信息检索研究与发展会议论文集（*SIGIR '23*），*2023*年*7*月*23*日至*27*日，中国台湾台北。ACM，纽约州纽约，美国，10页。https://doi.org/10.1145/3539618.3591878

1 引言

对话式推荐系统（CRSs）旨在捕捉用户在对话中表达的信息需求和偏好，并据此向用户推荐合适的商品 [12]。为促进对话式推荐系统的研究，大多数既有工作通过以下方式构建数据集：（i）根据现有用户行为合成对话 [5, 39]；（ii）在预先定义的交互模式 [10, 14, 16, 21, 22]下，通过众包模拟对话 fi。基于合成和模拟的对话，既有工作在开发对话式推荐系统方面取得了显著进展 [8, 11, 16, 29, 33, 39, 41]。然而，相关研究

*Weinan Zhang 是通讯作者。

允许免费为个人或课堂使用而制作本作品的全部或部分数字或纸质副本，前提是这些副本不得用于营利或商业用途，且副本首页须附有本声明及完整引文。本作品中由作者以外的其他方拥有版权的组件，必须尊重其版权。注明出处后，允许摘录。若要以其他方式复制、再版、发布至服务器或分发至列表，必须事先获得特定许可和/或支付费用。请向 permissions@acm.org 申请许可。
SIGIR '23，*2023*年*7*月*23*日至*27*日，中国台湾台北© 2023 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-9408-6/23/07...$15.00 https://doi.org/10.1145/3539618.3591878

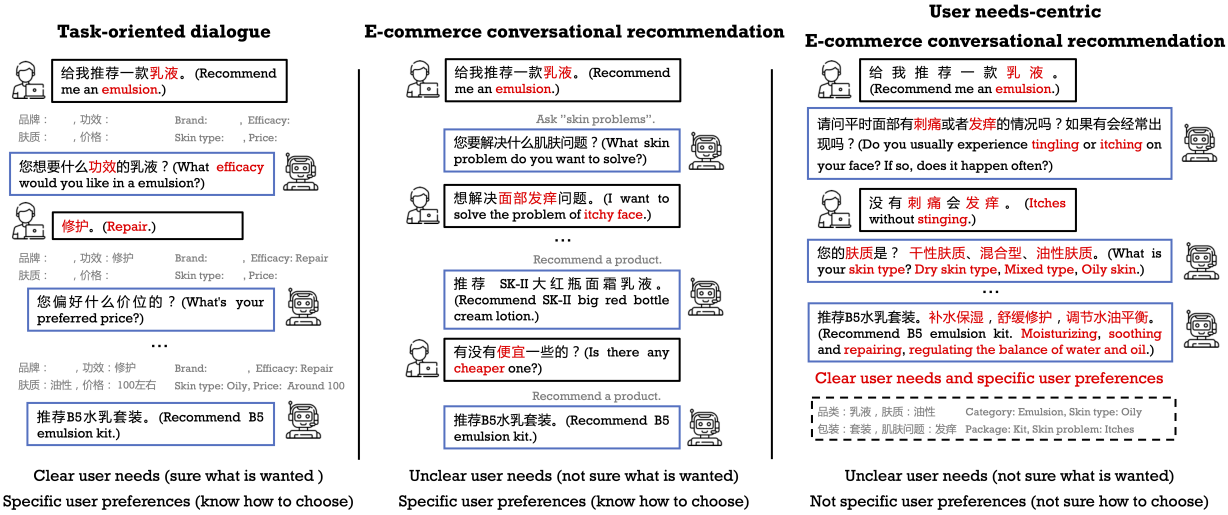

图1：任务导向型对话（TOD）、电子商务对话式推荐（ECR）与以用户需求为中心的电子商务对话式推荐（UNECR）的比较。TOD 以要素为对应的槽值对的槽值取决。主要关注点是获取用户偏好并推荐推荐，ECR 与 TOD 类似，但 ECR 还需要处理与对话上下文相关的复杂语义。UNECR 的主要关注点。

现有的对话推荐系统研究大多集中在电子商务场景，并利用众包工人与真实客服人员之间的对话作为数据集。例如，ReDial [21] 是一个基于电影领域的对话推荐数据集，其中两个众包工人根据预定义的电影偏好进行对话，以寻求或提供电影推荐。TG-ReDial [36] 通过将对话转换为主题引导的对话流程，扩展了 ReDial，以支持研究主题引导的对话推荐。DuRecDial [20] 是一个多领域对话推荐数据集，其中两个众包工人根据预定义的个人资料和目标进行对话。INSPIRED [6] 是一个人与人之间的对话推荐数据集，其中包含了对话中的社交策略注释。DuRecDial 2.0 [19] 是一个双语对话推荐数据集，其中两个众包工人根据预定义的个人资料和目标进行对话，并且每个对话都有两个平行版本，分别用中文和英文表达。

然而，这些现有的对话推荐数据集都是在实验室环境中构建的，其中用户和系统的角色由众包工人扮演，而不是真实的用户和商业推荐系统。因此，这些数据集中的对话与现实世界中的电子商务对话推荐场景存在较大差距。为了弥补这一差距，我们提出了 E-ConvRec，这是第一个基于电子商务客服对话的对话推荐数据集。E-ConvRec 中的对话是真实用户与客服人员之间的对话，涵盖了售前咨询、售后问题和闲聊等多种类型。因此，E-ConvRec 中的对话场景更加复杂和多样，对对话推荐系统提出了更大的挑战。

与现有的对话推荐数据集相比，E-ConvRec 还提供了更细粒度的标注，以支持对话推荐系统的开发和评估。具体来说，E-ConvRec 为每个对话提供了以下标注：(i) 对话中每个话语的意图；(ii) 对话中每个话语的槽值对；(iii) 对话中每个话语的推荐目标；(iv) 对话中每个话语的用户偏好；(v) 对话中每个话语的用户需求。这些标注可以帮助研究人员更好地理解对话推荐系统的工作原理，并开发出更有效的对话推荐系统。

用户需求。在售前对话中，用户会描述自己想要购买的产品特征，我们将其称为用户需求。我们以真实用户需求而非模拟用户需求作为线索，这在面向客服的电子商务对话式推荐（UNECR）中更为现实。UNECR 与 ECR 的一个关键区别在于用户需求的来源。在 ECR 中，用户需求通常是预定义的，并且在对话过程中保持不变。相比之下，UNECR 中的用户需求是动态的，并且可能会随着对话的进行而发生变化。例如，用户可能一开始只表达了一个模糊的需求，然后在对话过程中逐渐明确自己的需求。

任务。我们提出了售前对话中的三个子任务：(i) 对话理解任务，旨在理解用户在对话中表达的用户需求；(ii) 对话推荐任务，旨在根据用户需求推荐合适的产品；(iii) 对话生成任务，旨在生成包含能够引导用户需求的属性或满足用户需求的产品的回复；(iv) 对话评估任务，旨在评估对话系统在帮助用户明确自己的需求方面的有效性。

资源。我们收集了一个以用户需求为中心的电子商务对话推荐数据集（UNECR 数据集），UNECR 数据集包含 4,000 个对话和 91,000 个话语，来自一家领先的中国电子商务平台的真实用户与客服人员之间的对话。这些对话涉及 1,076 个产品和 3,812 个属性。对于售前对话中的每个话语，我们聘请专业的众包平台标注员为其标注了以下信息：(i) 话语的意图；(ii) 话语中提及的槽值对；(iii) 话语中推荐的目标产品；(iv) 话语中表达的用户偏好；(v) 话语中表达的用户需求。

基准。我们为这三个子任务建立了基准方法，并报告了这些方法在 UNECR 数据集上的性能。此外，我们还对这三个子任务的结果进行了分析，并发现：(i) 对于售前对话理解任务，所有基线方法在手机类别中的表现都低于其他类别。(ii) 对于对话推荐任务，所有基线方法在手机类别中的表现都低于其他类别。(iii) 对于对话生成任务，所有基线方法在手机类别中的表现都低于其他类别。我们还发现，UNECR 数据集中的手机类别与其他类别不同。

表 1：U-NEED 与现有相关数据集的比较。E-commerce(n) 表示该数据集涉及 n 个类别。Natural(c) 和 Natural(f) 分别指对话中关于用户需求的粗粒度和 fi 粒度标注。对于外部资源，UP 和 KB 分别指用户画像和知识库。UB 指对话前后的用户行为，而 UB(h) 指历史用户行为，即对话前的用户行为。Und.、Eli.、Rec.、Gen. 和 Eval. 分别指重点任务的五个方面，即理解、需求引导、推荐、生成和评估。

	统计信息					重点任务				
数据集	#对话	#话语	领域	对话来源	外部资源	Und.	Eli.	Rec.	Gen.	评估
FacebookRec [5]	1M	6M	电影	合成	-			✓		
ReDial [16]	10,006	182,150	电影	模拟		✓		✓	✓	
CCPE-M [28]	502	11,972	电影	模拟	-		✓			
GoRecDial [14]	9,125	170,904	电影	模拟	用户行为(h)			✓	✓	
INSPIRED [10]	1,001	35,811	电影	模拟	-			✓	✓	
TG-ReDial [39]	10,000	129,392	电影	合成	UP		✓	✓	✓	
MGConvRec [35]	7,615	73,971	餐馆	模拟	用户行为(h)	✓	✓	✓		
OpenDialKG [24]	15,673	91,209	电影，书籍	模拟	KB			✓		
DuRecDial [22]	10,190	155,447	电影，餐馆	模拟	用户画像，知识库		✓	✓	✓	
DuRecDial 2.0 [21]	10,190	255,346	电影，餐馆	模拟	用户画像，知识库		✓	✓	✓	
HOOPS[7]	-	11.6M	电子商务(4)	合成	KB		✓	✓		
E-ConvRec [13]	25440	775,338	电子商务	粗粒度(c)	用户画像, 知识库	✓		✓		
U-NEED（我们的方法）	7,698	53,712	**电子商务(5)**	**粗粒度(f)**	**用户行为**，知识库	✓	✓	✓	✓	✓

贡献。本文的主要贡献如下：

- 我们贡献了 U-NEED，这是一个源自电子商务场景的数据集。U-NEED 包含 7,698 fi个细粒度对话、333,879 条用户行为和 332,148 个产品知识元组。
- 为促进对售前对话中用户需求的研究，我们设计了 5 项任务（3 项针对售前对话，2 项针对用户需求）。据我们所知，在为对话式推荐系统构建资源时，我们率先设计了评测任务。 fi
- 我们为 5 项任务建立了基线，并报告了基准结果。 .
- 我们发布了实验所用的代码和脚本。

2 相关工作

相关工作涵盖两个方面：数据集和任务。我们在表 1 中详细比较了 U-NEED 与现有数据集。

2.1 对话式推荐数据集

为推动对话式推荐系统研究，围绕三个方向已有大量资源建设工作。

基于合成对话的数据集。一种简单的思路是结合用户行为、预先fi定义的交互策略和话语模板来合成对话。[1] Dodge 等人 [5] 使用 MovieLens 的评分和自然语言模板来合成一轮推荐。随后将其与一组问答结合，形成一个三轮对话，即 FacebookRec。Zhou 等人 [39] 提出了 TG-ReDial，使用豆瓣电影的评论。[2] TG-ReDial 的对话通过创建一个不断演进的主题线索来合成，该线索从先前主题引向推荐电影的目标主题。在 HOOPS 中，Fu 等人 [7]基于亚马逊评论构建了一个知识图谱 [25],然后提取关键实体以构建用户-物品交互，最后基于模板合成对话。通过合成对话，可以构建大量数据，用于训练对话式推荐系统（CRSs）的策略模块和推荐模块。由于话语是基于预定义且有限的模板构建的，因此很难利用合成数据训练对话式推荐系统生成多样化且具有说服力的回复。

基于模拟对话的数据集。大多数相关工作首先定fi义对话式推荐场景，然后通过众包模拟对话 [10, 14, 16, 21, 22, 24, 35]。其中，Li 等人 [16] 围绕电影推荐构建对话，即 ReDial，由一名工作人员扮演用户，另一名工作人员扮演推荐者。对话至少提及 4 部电影后结束。随后，一些工作为 ReDial 补充外部资源，例如维基百科 [1], 、ConceptNet [38] 和评论 [23]。Hayati 等人 [10] 遵循基于社会科学理论、与推荐策略相关的标注方案，构建了 INSPIRED。Liu 等人 [22] 通过要求推荐者主动引导对话，并结合求助者的兴趣进行推荐，模拟多类型对话，而不是由求助者向推荐者请求推荐。借助模拟对话，我们可以训练对话式推荐系统学习标注员提供的推荐理由。为便于收集数据，模拟场景中的交互模式往往简单且定义明确，因此，基于模拟数据集训练的对话式推荐系统只能适用于特定fi场景。此外，模拟对话的多样性和质量在一定程度上受限于标注员的习惯和知识背景。

基于自然对话的数据集。 合成对话和模拟对话都与真实场景中的对话存在较大差距。近期，Jia 等人 [13] 提出了一个数据集 E-ConvRec，其中包含用户与客服围绕售前主题展开的对话

[1]MovieLens 评分：https://grouplens.org/datasets/movielens/。[2] 豆瓣电影：https://movie.douban.com/。

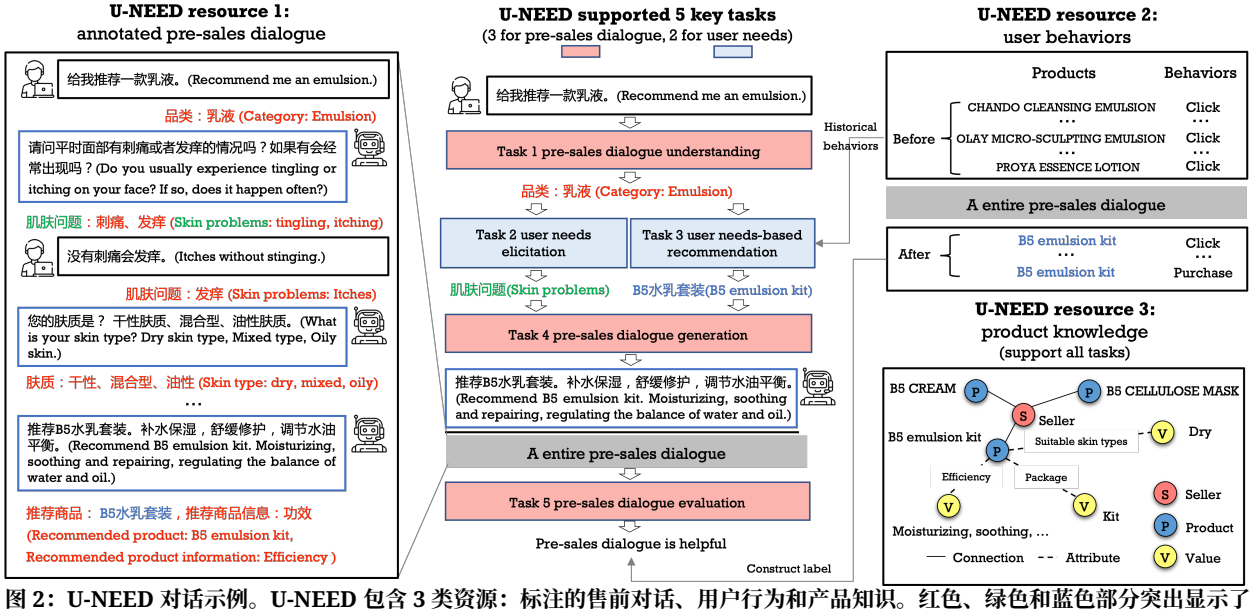

图 2：U-NEED 对话示例。U-NEED 包含 3 类资源：标注的售前对话、用户行为和产品知识。红色、绿色和蓝色部分突出显示了与用户需求相关的标注。U-NEED 支持两个方面的 5 项关键任务：售前对话和用户需求。售前对话的 3 项任务以红色方框表示，用户需求的 2 项任务以蓝色方框表示。中间部分还展示了 5 项任务之间的联系。灰色线条表示资源与任务之间的联系。

staff in 电子商务场景。与 E-ConvRec 相同，我们基于电子商务场景中的自然对话构建 U-NEED。

与现有数据集相比，U-NEED 具有以下优势：(i) U-NEED 提供了关于售前对话中用户需求的 fi粒度标注。借助 U-NEED，我们可以开发面向售前场景的电子商务对话式推荐系统。(ii) U-NEED 涵盖了 5 个热门类别，为分析用户在不同类别中的需求提供了线索。(iii) U-NEED 提供了对话前后的用户行为信息，使我们能够探索评估对话式推荐系统的方法。

2.2 对话式推荐中的任务

参考对话式推荐系统的一般框架 [8]，，我们从五个方面总结现有数据集重点关注的任务：理解、需求引导、推荐、生成和评估。

理解。 我们将理解任务视为从对话上下文中捕捉用户需求或偏好。在 ReDial 中，Li 等人 [16] 提出了一项任务，根据话语推断情感和观点。Xu 等人 [35] 设计了一项任务，将用户在对话中表达的偏好预测为观点目标（某个项目或值）及其对应情感极性（即积极、消极和中性）组成的配对。Jia 等人 [13] 提出了用户偏好识别任务，用于识别用户话语中的描述性词语、类别词、比较词和否定词。与以往工作不同，我们提出的理解任务不仅识别与需求或偏好相关的词语，还识别其中涉及的产品属性。

需求引导。 我们将需求引导任务视为识别能够引导用户表达需求或偏好的信息。请注意，何时提问或推荐产品的策略 [13, 14] 不属于需求引导的范畴。在 TG-ReDial [39] 和 DuRecDial [22],中，该任务旨在预测下一个主题或目标，以引导对话逐步接近推荐目标。在 HOOPS [7] 和 MGCon-vRec [35], 中，需求引导任务主要是选择一个询问用户对某个槽位或值的偏好的问题。与以往工作不同，我们提出的需求引导任务会选择多个能够引导用户表达需求的属性。

推荐。 我们将推荐任务视为推荐满足用户需求或偏好的商品。在大多数以往的工作中，推荐任务旨在根据对话上下文 [5, 7, 14, 16, 21, 22, 24, 35, 39] 推荐用户可能会接受的商品。其中，Jia 等人 [13]将个性化推荐设计为一个排序任务，根据用户资料、产品知识库（知识库）和对话上下文来判断用户是否会购买候选产品。我们的推荐任务定义为推荐满足所获取用户需求的产品。

生成。 我们将生成任务视为生成包含给定信息的回复。大多数先前工作旨在生成对话话语，以类似人类说话者 [14, 21, 22] 的方式与求助者交谈。此外，一些先前工作旨在将主题 [39] 或社交策略 [10] 融入回复中，以引导对话。Li 等人 [16] 将生成任务设计为生成包含推荐电影的回复。

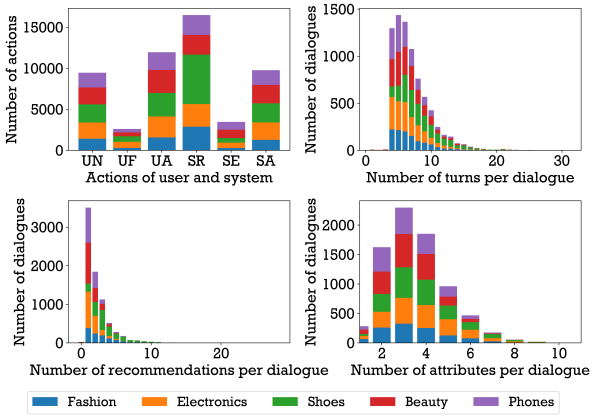

图3：标注的售前对话统计信息。UN、UF 和 UA 分别是用户需求、用户反馈和用户查询的缩写，SR、SE 和 SA 分别指推荐、解释和问题。

与现有工作不同，我们提出的生成任务旨在生成与需求引导任务识别出的属性或推荐任务预测出的商品相关的回复。

评测。 据我们所知，我们是第一个构建对话推荐系统（CRS）数据集来评测任务的工作。我们将任务视为评估用户与对话式推荐系统进行对话时是否有助于明确其购买需求。

与现有数据集相比，据我们所知，U-NEED 是首个满足以下条件的电子商务对话式推荐数据集：(i) 支持对话式推荐中的所有常见任务；(ii) 包含一项以往研究无法支持的评测任务。

3 数据集

我们构建了一个领先的中国电子商务平台[3]的电子商务对话推荐数据集 U-NEED。U-NEED 数据集包含 3 类资源：(i) 标注的售前对话；(ii) 用户档案；(iii) 产品知识。

3.1 标注的售前对话

3.1.1 对话收集。我们首先从人际对话中收集语料，然后进行细粒度的人工标注和脱敏处理以构建标注的售前对话。我们收集了2022年9月1日至2022年10月1日期间淘宝客服与客户之间的售前对话。为了涵盖不同类别的售前对话，我们收集了三个类别的对话，即 *鞋类*、*时尚* 和 *鞋类*。为了模拟真实的购物场景，我们雇佣了部署在线生产环境的聊天机器人来模拟与客户的对话。除了去除个人信息外，我们没有对对话做任何修改。最终，我们收集了 *Appen* 对售前对话。

[3]https://www.taobao.com/

表 2：用户行为元组的示例数据。

产品ID	用户ID	行为ID	时间戳
1	2	3	2022-05-05 16:25:25
1	2	5	2022-05-05 17:26:55

表3：产品知识元组的示例数据。

产品ID	属性	值	卖家ID
1	价格:Ù（价格范围）	-（中间）	2
1	üH（Effect）	防皱（抗皱）	2
1	¡L（品牌）	âp（Olay）	2

对话我们雇用了一名具有相关学科的员工来负责标注对话的动作及的属性和推荐产品。标注流程的详细信息如处。[5] 其余 名员工负责质量控制我们将 Appen 的 对话交给 Appen，随后收到了7,698条符合标准的售前对话。对于标注售前对话中的每条话语，我们支付 0.68元人民币的标注费用。

3.1.2 对话分析。我们收集了7,698份细粒度标注的售前对话，其中包括来自手机、时尚鞋类和电器类别的1662、1513、1135、1748和1640段对话。如图3所示，我们从四个关键方面展示了标注的售前对话统计数据。

用户和系统的动作。 图3左上部分展示了各类别标注的售前对话中用户和系统的动作的分布。纵轴表示对话中相应动作的出现次数。我们发现：i）在六种预定义动作中，SR出现次数最多，其次是UA和UF，而SE和SA出现次数最少。ii）各类别在同类动作的数量上达到相同。与 *鞋类* 和 *时尚* 类别相比，*鞋类* 类别中 SR 出现得更频繁。

轮次数量。 图3右上部分展示了不同类别中各种长度的对话数量。我们观察到：(i) 对话的轮次数量主要在5到10轮之间。这表明，在实际场景中，简短对话可能更为合适。(ii) 此外，随着轮次数量增加，*鞋类* 所占比例逐渐上升。这可能是因为很难描述与鞋类相关的需求和偏好，因此需要更多轮次。

推荐数量。 图3左下部分统计了标注的售前对话中的推荐数量。我们发现：(i) 仅推荐1种产品的对话占所有对话的三分之一以上。(ii) 然而，在 *鞋类* 类别中，对话包含更多推荐。

属性数量的数量。 图3右下部分统计了标注的售前对话中对话涉及的属性数量。我们发现：(i) 涉及1至3个属性的对话数量最多，而涉及5个属性的对话数量最少。(ii) 各类别在属性数量上比均衡。

[4]https://www.appen.com.cn/ [5] https://github.com/LeeeeoLiu/U-NEED/blob/main/annotation.md

3.2 相关资源

3.2.1 用户行为。对于每位参与标注的售前对话的用户，我们都会收集其行为，以分析售前对话是否对用户有所帮助。我们使用元组 $(p,\ u,\ b,\ t)$ 表示用户 u 在时间 t 对产品 p 执行了行为 b。表 2 展示了 2 个用户行为元组的示例数据。每行行为数据都包含用户ID、产品ID、行为ID 和行为的时间戳。行为包括点击、购买、收藏和加入购物车。实际操作中，我们会收集用户与客服人员进行售前对话当天的用户行为。随后，我们会对收集到的行为进行严格的脱敏操作，使其无法被追溯。

我们收集了333,879条用户行为元组，涉及7,620名用户和68,027种产品。根据所收集对话的时间戳，我们可以将用户的行为分为两部分：对话开始前的用户行为称为用户历史行为；对话开始后的用户行为则视为用户后续行为。借助用户历史行为和后续行为，我们可以构建一个标签来评估对话式推荐系统。

3.2.2 产品知识。我们收集了标注的售前对话中涉及的产品信息。我们使用一个元组（$p,\ a,\ v,\ s$）表示卖家 s 拥有产品 p，且产品 p 的属性 a 的值为 v。表3展示了3个产品知识元组的示例。每一行由产品ID、属性、值和卖家ID组成。

我们收集了332,148个产品知识元组，涉及68,954个产品和13,122个卖家。借助这些产品知识，我们可以构建异构图 [26]，以支持未来对售前对话的研究。

3.3 研究主题

基于对U-NEED及其提供资源的深入分析，我们概述了5个有前景的、以用户需求为中心的电子商务对话式推荐（UNECR）研究主题（RT）。

RT1：理解话语中表达的用户需求。真实场景与模拟场景之间的一个显著差异在于用户群体。U-NEED包含大量用户对其需求的多样化描述。在手机类别中，表达功能需求的方式超过370种。

RT2：识别诱发用户需求的属性。在真实场景中，用户的需求通常较为模糊，也不熟悉产品属性。有经验的客服人员可以迅速引导用户明确其需求。U-NEED提供了每轮对话中客服人员询问的属性标签，为研究用户需求与产品属性之间的关系提供了线索。

RT3：推荐满足用户需求的产品。在真实场景中，推荐更具挑战性。大多数用户的耐心有限，客服人员ff必须在对用户需求或偏好知之甚少的情况下，推荐合适的产品。借助U-NEED，我们可以探索客服人员ff在真实场景中提供的推荐。

RT4：生成包含能够引出用户需求的属性，或满足用户需求的商品的回复。对话式推荐系统需要将给定信息转化为用户容易理解的形式。例如，客服人员ff 询问用户对“皮肤问题”的偏好时，会将其表述为“是刺痛还是发痒？”。

RT5：评估售前对话是否有助于明确用户需求。U-NEED包含售前对话前后的用户行为，使我们能够构建用于评估对话式推荐系统的标签，并探索售前对话的影响。

4 项任务

为促进对这 5 个主题的研究，我们设计了 5 项关键任务。

4.1 售前对话理解（任务 1）

针对 RT1，我们设计了一项售前对话理解任务。该任务旨在通过以下方式理解话语：（i）识别与用户需求相关的属性；（ii）提取相应的偏好。

给定一个话语 x，例如“有补水和保湿套装吗？”，模型 M_u 旨在预测语义框架 $U=\{(a_i,\ v_i)\}$，即 { (“功效”, “补水和保湿”), (“包装”, “套装”) }，这些框架包含在 x 中。形式化地，该任务表示为：

$$U = \mathcal{M}_u(x). \tag{1}$$

4.2 用户需求 elicitation（任务 2）

对于 RT2，我们设计了一项用户需求 elicitation 任务，旨在选择能够引出更多用户需求信息的属性。

给定售前对话上下文 $X=\{x_i\}$ 和 X 的上下文语义框架 $C=\{U_i\}$，模型 M_a 旨在选择能够引出用户需求信息的属性 $A=\{a_i\}_{i=1}$。形式上，用户需求 elicitation 任务表示为：

$$A = \mathcal{M}_a(X, C). \tag{2}$$

4.3 基于用户需求的推荐（任务 3）

对于 RT3，我们设计了一项基于用户需求的产品推荐任务，旨在推荐满足售前对话中表达的显性用户需求和隐性用户需求（即用户在售前对话之前的行为）的产品。

给定售前对话上下文 $X=\{x_i\}$ 和用户历史行为 $B=\{(p_i,\ b_i)\}$，模型 M_r 旨在推荐同时满足显性和隐性用户需求的产品 $P=\{p_i\}$。形式上，该任务表示为：

$$P = \mathcal{M}_r(X, B). \tag{3}$$

4.4 售前对话生成（任务4）

对于 RT4，我们设计了一项售前对话生成任务，旨在生成包含给定用户需求信息的售前对话回复。

给定售前对话上下文 $X=\{x_i\}$、满足用户需求的产品 $P=\{p_i\}$，以及能够引出用户需求的属性 $A=\{a_i\}$，模型 M_g 旨在生成回复 $Y=\{w_i\}$。形式化地，该任务表示为：

$$Y = \mathcal{M}_g(X, A, P). \tag{4}$$

A 或 P 可以是 ;，这意味着没有合适的属性/产品。

4.5 售前对话评估（任务5）

对于 RT5，我们设计了一项售前对话评估任务，旨在衡量售前对话是否有帮助。

表4：基线方法在3个典型类别的5项任务上的表现：美妆、时尚和鞋类。**为便于呈现结果，我们对部分评估指标进行了缩写：P和R分别表示精确率和召回率。H@K和M@K分别表示Hit@K和MRR@K。Info.和Rel.分别表示信息量和相关性。** κ **表示标注员之间两两Cohen's kappa系数的平均值。PCC、SCC和余弦相似度分别表示皮尔逊相关系数、斯皮尔曼相关系数和余弦相似度。最佳结果以粗体突出显示。**

	美妆			鞋类			手机			全部 5 个类别		
	任务1：售前对话理解											
方法	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Bert [4]	0.5355	0.6284	0.5782	0.5851	0.7020	0.6382	0.4212	0.5384	0.4726	0.4549	0.5652	0.5041
Bert+条件随机场 [31]	0.6731	0.6802	0.6766	0.7302	0.7703	0.7497	0.5620	0.5923	0.5768	0.6688	0.6530	0.6608
Bert+双向长短期记忆网络+条件随机场 [3]	**0.7282**	**0.7481**	**0.7380**	**0.7870**	**0.8101**	**0.7984**	**0.6701**	**0.6990**	**0.6843**	**0.6892**	**0.6875**	**0.6884**
	任务 2：用户需求 elicitation											
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DiaMultiClass [18]	0.4037	**0.7228**	**0.5054**	0.3361	**0.4131**	0.3423	**0.4534**	**0.5212**	**0.4585**	0.3222	**0.4966**	**0.3662**
DiaSeq [18]	**0.4761**	0.4272	0.4424	**0.3992**	0.3305	**0.3498**	0.4414	0.3789	0.3966	**0.3555**	0.2996	0.3153
	任务 3：基于用户需求的推荐											
	H@10	H@50	M@50	H@10	H@50	M@50	H@10	H@50	M@50	H@10	H@50	M@50
Bert [4]	**0.2985**	**0.3938**	**0.1532**	**0.1014**	**0.2550**	**0.0384**	0.4275	0.7174	0.2086	**0.1593**	**0.3310**	**0.0660**
SASRec [15]	0.1477	0.3046	0.0378	0.0722	0.1720	0.0344	0.4022	**0.7246**	0.1695	0.1400	0.2747	0.0585
TG-CRS [39]	0.1754	0.2585	0.0778	0.0568	0.1521	0.0231	**0.4565**	0.6920	**0.2145**	0.1470	0.2500	0.0655
	任务 4：售前对话生成											
	Dist-4	信息。	Rel.	Dist-4	信息。	Rel.	Dist-4	信息。	Rel.	Dist-4	信息。	Rel.
Transformer [34]	0.3714	**1.2767**	0.5267	0.3718	1.0400	0.9600	**0.4037**	**1.2267**	0.9467	0.2806	**1.1567**	0.8800
GPT-2 [27]	**0.3811**	0.6433	0.2767	0.3357	0.5000	0.3767	0.3266	0.7633	0.4700	**0.2905**	0.5700	0.4267
KBRD [1]	0.2808	1.2233	0.6133	0.4051	**1.0933**	1.0467	0.4157	1.0900	0.9867	0.2233	1.1367	0.9167
NTRD [19]	0.3275	1.1500	**0.6867**	**0.4635**	1.0100	**1.0933**	0.3775	1.0400	**1.0567**	0.2489	1.0033	**0.9900**
κ	-	0.7162	0.3947	-	0.8431	0.6503	-	0.6862	0.5759	-	0.7523	0.4958
	任务5：售前对话评估											
	PCC	SCC	Cos.	PCC	SCC	Cos.	PCC	SCC	Cos.	PCC	SCC	Cos.
DEB [30]	**0.1642**	**0.1628**	**0.9327**	**0.1504**	**0.1963**	0.9097	**0.2678**	**0.2815**	**0.9366**	**0.1617**	**0.1864**	0.9212
P值	<0.0299	<0.0313	-	<0.0416	<0.0076	-	<0.0015	<0.0008	-	<6e-06	<1e-07	-
Bert-RUBER [9]	0.0901	0.1133	0.9218	0.0916	0.1157	**0.9219**	0.0900	0.1141	0.9218	0.0742	0.1092	**0.9214**
P值	<0.0126	<0.0017	-	<0.0111	<0.0013	-	<0.0126	<0.0015	-	<0.0398	<0.0024	-

给定一段完整的售前对话 $D = \{x_i\}$，模型 M_e 旨在预测售前对话 D的评分 r。形式上，售前对话评估任务表示为：

$$r = \mathcal{M}_e(D). \tag{5}$$

5 实验与结果

表4展示了5项任务的基准结果。以下小节将分别介绍每项任务的数据准备、基线方法、评估指标以及结果与分析。

随后，我们使用标注的售前对话的时间戳，将整个数据集划分为训练集、验证集和测试集。U-NEED 涵盖 30 天的数据，我们将前 24 天的对话用作训练集，最后 3 天的对话用作测试集，其余对话则用作验证集。随后，我们构建每项任务所需的输入和标签。为了便于复现，关于实验结果，我们公开了数据预处理和实验中使用的代码及脚本。[6]

5.1 任务1基准

5.1.1 数据准备。基于人工标注的属性和偏好，我们使用 BIO 标注为话语中的每个词构建标签。[7] 对于话语"有补水和保湿套装吗？"及其标注的属性和偏好 { ("Efficacy", "补水和保湿"), ("包装", "套装") }，构建的标签如下："Are/*O* there/*O* any/*O*hydration/*B*-美妆-*Efficacy* and/*I*-美妆-*Efficacy* moisturizing/*I*- 美妆-*Efficacy* kits/*B*-美妆-包装 ?/*O*"。我们构建了41,592个训练样本、6,251个验证样本和2,874个测试样本。

5.1.2 基线方法。我们选择了3种在电子商务场景的生产环境中具有良好效果并得到应用的模型。

[6] https://github.com/LeeeeoLiu/U-NEED [7] https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging)

Bert [4] 根据输入话语的表示预测每个标签的概率。

Bert+条件随机场 [31] 在考虑输入话语的表示以及预测标签之间的序列关系后，预测每个标签的概率。

Bert+双向长短期记忆网络+条件随机场 [3] 在 Bert 编码后考虑双向信息，以获得输入话语更好的表示。它还考虑了预测标签之间的序列关系。

5.1.3 评估指标。 我们采用精确率、召回率和F1值作为评估指标。

精确率 是正确识别的标签数量与所有识别出的标签数量之比。

召回率 是正确识别标签的数量与原本正确标签数量的比率。

F1值 的计算方式是精确率和召回率的平均值。

5.1.4 结果与分析。 从表4可以看出，在全部5个类别中，Bert+双向长短期记忆网络+条件随机场模型取得了最佳结果。Bert+双向长短期记忆网络+条件随机场模型在鞋类 类别中的F1值为0.7984，而在手机 类别中，F1值仅为0.6843。通过分析话语，我们发现，用户对鞋类的需求表达在不同对话中相对相似，而手机 类别中的差异则很大。对用户不规则表达与细粒度产品属性之间的关系进行建模，可能是未来研究中值得关注的主题。

5.2 任务2基准测试

5.2.1 数据准备。 我们根据客服人员询问属性偏好的话语构建样本。以图2中标注的售前对话为例。在对话的第二轮中，系统向用户询问有关“皮肤问题”的问题。输入是对话上下文和识别出的用户需求，即“类别：乳液”。标签是客服在这一轮询问的属性，即“皮肤问题”。请注意，标签可能包含多个属性。例如，“您对洗衣机的功能有什么要求？预算是多少？”同时包含“功能”和“价格”。我们构建了7,432个训练样本、1,108个验证样本和1,042个测试样本。

5.2.2 基线方法。 受 [18], 启发，我们从多标签分类和序列生成中设置了2个基线模型。
DiaMultiClass [18] 对输入文本进行编码，然后预测每个属性的概率。在推理阶段，我们将阈值设为0.5，以输出预测结果。

DiaSeq [18] 利用基于GRU的解码器来预测属性组合。

5.2.3 评估指标。
精确率 是正确选出的属性数量与所有选出的属性数量之比。
召回率 是正确选出的属性数量与原本正确的属性数量之比。
F1值 通过计算精确率和召回率的平均值之和得出。

5.2.4 结果与分析。 从表4可以看出，DiaMultiClass 在所有类别上的召回率均优于 DiaSeq。此外，我们在美妆 和手机 类别中观察到一种现象：DiaMultiClass 的精确率不如 DiaSeq，但召回率和 F1值均优于 DiaSeq。对于鞋类，我们看到两种方法在 F1值方面取得了相当的结果。

5.3 任务3基准测试

5.3.1 数据准备。 我们根据客服人员在售前对话中推荐产品的话语构建样本。以图2中的标注售前对话为例。在对话的最后一轮，系统推荐了一款产品“B5乳液套装”。输入是对话上下文和用户的历史行为。标签是本轮推荐的产品，即“B5乳液套装”。请注意，客服人员可能一次推荐多个产品。按照 [37, 38, 40], , 我们将多个推荐作为输入相同的多个样本进行处理。我们构建了13,671个训练样本、2,064个验证样本和1,864个测试样本。

5.3.2 基线方法。我们考虑了 3 种广泛应用的方法，它们使用不同类型的信息进行推荐。
Bert [4] 对对话上下文进行编码，然后生成推荐。**SASRec [15]** 采用自注意力机制，捕捉用户历史行为中的长期语义信息，从而生成推荐。**TG-CRS [39]** 融合对话上下文和用户历史行为的表示，以生成推荐。

这里我们通过使用 Bert 模型对对话上下文进行编码，并将用户历史行为作为主题线程，将 TG-CRS 修改为 fit 以适应我们的任务。

5.3.3 评估指标。 我们采用两种广泛使用的指标来评估前K(K = 1，10，50) 项推荐性能 [17, 40]。**Hit@**K 表示在所有相关项目中，排名前K 位返回的相关项目所占的比例。

MRR@K 计算方法是：对排名前K 位中返回的项目的倒数排名取平均值。如果某个项目未出现在排名前K 位中，则其倒数排名为 0。

5.3.4 结果与分析。 从表4可以看出，不同类别的推荐效果存在较大差距。在手机 类别中，各模型的表现均显著优于其他类别。具体而言，TG-CRS 在手机 类别中表现出色，在 H@10 和 M@50 上均取得最佳结果。在其他三个类别中，Bert 在所有指标上均取得最佳结果。

5.4 任务4的基准

5.4.1 数据准备。 我们使用售前对话中客服人员的回复构造样本。输入为售前对话上下文，真实值为客服人员的回复。我们构建了 15,339 个训练样本、1,958 个验证样本和 1,932 个测试样本。

5.4.2 基线方法。我们设置了4种基线方法。其中2种广泛用于回复生成，另外2种是生成包含对话式推荐系统中推荐项目的回复的方法。

Transformer [34] 采用基于 Transformer 的编码器-解码器框架来生成恰当的回复。**GPT-2 [27]** 是一种预训练文本生成模型，并在 U-NEED 数据集上进行 fi微调。**KBRD [1]** 采用 Transformer，并基于知识图谱增强词语权重建模。**NTRD [19]** 先生成带有特殊标记的回复模板，再将特殊标记替换为推荐商品。

5.4.3 评估指标。我们采用 3 种常用指标来评估生成表现 [40]。

Distinct@4 的计算方式是：计算回复中不同 4-gram 数量占所有 4-gram 数量的比例，并取其平均值，以此衡量生成回复的多样性。**信息量**通过计算所有生成回复的信息量程度的平均值来衡量。分数范围为 {0，1，2}，我们将所有标注员给出的分数取平均，作为fi最终分数。

相关性是所有生成回复相关度的平均值。标注员对生成的回复包含与真实值中相关的产品信息的程度进行评分。

信息性和相关性用于人工评估，我们招募了 12 名标注员，对在 美妆、鞋类、手机 以及全部 5 个类别中随机选取的 100 条生成回复进行评估。得分为 {0,1,2}，我们将所有标注员的得分取平均值，作为最终得分。我们计算 Fleiss's kappa [6] 以衡量标注员之间的一致性。

5.4.4 结果与分析。从表4中可以看出，在所有类别中，KBRD 和 NTRD 在相关性方面均优于 Transformer 和 GPT-2。这是因为前者在生成回复的过程中显式地融入了产品信息。此外，我们观察到，美妆 类别中的相关性得分远低于其他类别，而且标注员似乎有不同的判断。这是因为美妆 类别中的大多数售前对话都围绕产品功效展开，通常涉及许多非标准的需求描述，这使得美妆 类别中的售前对话生成颇具挑战性。

5.5 任务5的基准

5.5.1 数据准备。受 [2], 的启发，我们根据用户在售前对话中推荐的产品与用户后续行为之间的相关性，构建客观评分。我们构建了 5,517 个训练样本、821 个验证样本和 767 个测试样本。

5.5.2 基线方法。我们选择了 2 种对话评估中常用的基线方法。

DEB [30] 在 7.27 亿条 Reddit 对话上进行预训练，然后在 DailyDialog++ 数据集上进行微调。
Bert-RUBER [9] 对广泛应用的评估模型 RUBER [32]进行了三项改动：将 word2vec 嵌入替换为 bert 嵌入，将双向循环神经网络替换为池化策略，并将排序损失替换为分类损失。
我们使用售前对话对 Bert 模型进行 fi微调，用于 DEB 和 Bert-RUBER。

5.5.3 评估指标。 接下来 [9], 我们采用 3 种常用指标来评估基线模型的性能。**皮尔逊** 指皮尔逊相关系ffi数，用于衡量-

衡量两个有序变量之间的线性相关性。
斯皮尔曼 指斯皮尔曼相关系ffi数，用于衡量任意单调关系。

余弦相似度 用于计算模型生成的分数与根据用户行为构建的分数之间的相似程度。

5.5.4 结果与分析。从表4可以看出，DEB 在所有相关性系数指标上均优于 bert-RUBER。DEB 在手机类别中的表现优于其他类别。这可能是因为手机类别中产品优劣的评判标准更明确，例如屏幕尺寸较大、特定的摄影像素值较高。而鞋类和美妆的评判标准则更复杂，涉及产品外观和个人需求。

6 数据集访问

我们遵循严格的政策和规则，在收集和处理售前对话时保护用户的隐私。由于 U-NEED 收集自真实场景，访问 U-NEED 需要fi填写表单。[8] 申请获批后，我们会通过电子邮件向您发送 U-NEED。请注意，U-NEED 仅供研究用途。未经许可，不得将其用于任何商业用途，也不得分发给他人。我们提供 5 个公开可用的示例数据。[9] 这些示例经过适当修fi改，以展示售前对话的特征，但不包含事实信息。

7 结论

本文聚焦于真实用户需求在售前对话中的主导作用，并将其称为 UNECR。为促进 UNECR 研究，我们提出了一个 fi粒度的数据集 U-NEED，其中包含标注的售前对话、用户行为和产品知识元组，涵盖 5 个顶级类别。基于 U-NEED，我们提出了 5 项关键任务，并为售前对话和用户需求建立了 15 种基线方法。根据 U-NEED 上 5 项任务的实验结果与分析，我们发现：(i) 用户在不同类别中描述其需求的方式存在很大差异。在时尚 类别中，用户需求的描述比鞋类 类别更加多样。(ii) 不同类别中的推荐数量有所不同。在鞋类类别中，售前对话中的推荐数量多于其他类别。

U-NEED 在三个方面具有巨大价值：新的意义、实际应用以及未来研究。(i) U-NEED 表明，用户需求是探索电子商务场景下对话式推荐任务的重要且有效的切入点。(ii) 借助

[8]https://github.com/LeeeeoLiu/U-NEED/blob/main/dataset_access.md [9] https://github.com/LeeeeoLiu/U-NEED/blob/main/examples.md

通过 U-NEED，我们能够开发可应用于真实场景的对话式推荐系统（CRSs）。(iii) U-NEED 支持未来关于售前对话和用户需求的研究。

致谢

我们感谢匿名审稿人提出的宝贵意见。本工作得到中国科技创新2030重大项目（编号：2020AAA0108605）和国家自然科学基金（编号：62076081、61772153和61936010）的支持。

参考文献

[1] Qibin Chen、Junyang Lin、Yichang Zhang、Ming Ding、Yukuo Cen、Hongxia Yang 和 Jie Tang。2019。基于知识的对话推荐系统。载于 *EMNLP* 论文集。[2] Yinpei Dai、Wanwei He、Bowen Li、Yuchuan Wu、Zheng Cao、Zhongqi An、Jian Sun 和 Yongbin Li。2022。CGoDial：用于中文任务导向型对话评估的大规模基准。载于 *2022* 年自然语言处理经验方法会议论文集。计算语言学协会，阿布扎比，阿拉伯联合酋长国，4097–4111。https: //aclanthology.org/2022 274 .emnlp-main。Zhenjin Dai、Xutao [3] Wang、Pin Ni、Yuming Li、Gangmin Li 和 Xuming Bai。2019。使用 BERT 双向长短期记忆网络条件随机场进行中文电子健康记录中的命名实体识别。载于 *2019* 年第 *12* 届图像、信号处理、生物医学工程与信息学国际大会，*CISP-BMEI 2019*。[4] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。用于语言理解的深度双向变换器预训练。载于*2019* 年计算语言学协会北美分会会议论文集：人类语言技术，第 *1* 卷（长文与短文）。计算语言学协会，明尼阿波利斯，明尼苏达州，4171 4186 10 18653 1 19 1423 –。https://doi.org/。/v /N - [5] Jesse Dodge、Andreea Gane、Xiang Zhang、Antoine Bordes、Sumit Chopra、Alexan-der H. Miller、Arthur Szlam 和 Jason Weston。2016。端到端对话系统学习的先决质量评估。载于 *ICLR* 论文集。[6] Joseph L Fleiss。1971。多名评定者之间的名义尺度一致性测量。《心理学公报》（1971）。[7] Zuohui Fu、Yikun Xian、Yaxin Zhu、Shuyuan Xu、Zelong Li、Gerard de Melo 和 Yongfeng Zhang。2021。HOOPS：人在回路图推理，用于对话式推荐。载于 *SIGIR* 论文集。[8] Chongming Gao、Wenqiang Lei、Xiangnan He、Maarten de Rijke 和 Tat-Seng Chua。2021。对话推荐系统的进展与挑战：综述。*AI Open*（2021）。[9] Sarik Ghazarian、Johnny Wei、Aram Galstyan 和 Nanyun Peng。2019。利用上下文化嵌入改进开放域对话系统的自动评估。载于神经语言生成优化与评估方法研讨会论文集。[10] Shirley Anugrah Hayati、Dongyeop Kang、Qingxiaoyang Zhu、Weiyan Shi 和 Zhou Yu。2020。INSPIRED：迈向社交型推荐对话系统。载于 *EMNLP* 论文集。[11] Chenhao Hu、Shuhua Huang、Yansen Zhang 和 Yubao Liu。2022。学习推断对话式推荐中的用户隐式偏好。载于 *SIGIR* 论文集。[12] Dietmar Jannach、Ahtsham Manzoor、Wanling Cai 和 Li Chen。2021。对话推荐系统综述。*ACM* 计算评论 54，5，文章 105（2021 年 5 月），36 页。https://doi.org/10.1145/3453154[13] Meihuizi Jia、Ruixue Liu、Peiying Wang、Yang Song、Zexi Xi、Haobin Li、Xin Shen、Meng Chen、Jinhui Pang 和 Xiaodong He。2022。{E}-{C}onv{R}ec：面向电子商务客服的大规模对话推荐数据集。载于 第十三届语言资源与评估会议论文集[14] Dongyeop Kang、Anusha Balakrishnan、Pararth Shah、Paul Crook、Y-Lan Boureau 和 Jason Weston。2019。作为沟通博弈的推荐：面向目标对话的自监督机器人对战。载于 *EMNLP* 论文集。[15] Wang-Cheng Kang 和 Julian McAuley。2018。自注意力序列推荐。载于 *2018 IEEE* 国际数据挖掘会议（*ICDM*）。197–206。https://doi.org/10.1109/ICDM.2018.00035[16] Raymond Li、Samira Ebrahimi Kahou、Hannes Schulz、Vincent Michalski、Laurent Charlin 和 Chris Pal。2018。迈向深度对话推荐。载于 *NeurIPS* 论文集。[17] Shuokai Li、Ruobing Xie、Yongchun Zhu、Fuzhen Zhuang、Zhenwei Tang、Wayne Xin Zhao 和 Qing He。2022。对话中的自监督学习推荐。信息处理与管理（2022）。[18] Ziming Li、Julia Kiseleva 和 Maarten de Rijke。2020。重新思考面向任务的对话系统中的监督学习与强化学习。载于 EMNLP Findings 论文集。[19] Zujie Liang、Huang Hu、Can Xu、Jian Miao、Yingying He、Yining Chen、Xiubo Geng、Fan Liang 和 Daxin Jiang。2021。学习对话推荐系统的神经模板。载于 EMNLP 论文集。[20] Yuanxing Liu、Zhaochun Ren、Wei-Nan Zhang、Wanxiang Che、Ting Liu 和 Dawei Yin。2020。关键词生成改进电子商务基于会话的推荐。载于 WWW 论文集。[21] Zeming Liu、Haifeng Wang、Zheng-Yu Niu、Hua Wu 和 Wanxiang Che。2021。DuRecDial 2.0：用于对话式推荐的双语平行语料库。载于 EMNLP 论文集。[22] Zeming Liu、Haifeng Wang、Zheng-Yu Niu、Hua Wu、Wanxiang Che 和 Ting Liu。2020。迈向多类型对话中的对话式推荐。载于 ACL 论文集。[23] Yu Lu、Junwei Bao、Yan Song、Zichen Ma、Shuguang Cui、Youzheng Wu 和 Xi-aodong He。2021。RevCore：评论增强的对话式推荐。载于 ACL Findings 论文集。[24] Seungwhan Moon、Pararth Shah、Anuj Kumar 和 Rajen Subba。2019。OpenDi-alKG：利用知识图谱上的基于注意力的游走实现可解释的对话式推理。载于 ACL 论文集。[25] Jianmo Ni、Jiacheng Li 和 Julian McAuley。2019。利用远程标注评论和细粒度方面为推荐提供依据。载于 EMNLP 论文集。[26] Xichuan Niu、Bofang Li、Chenliang Li、Rong Xiao、Haochuan Sun、Hongbo Deng 和 Zhenzhong Chen。2020。用于改进电子商务商店搜索长尾性能的双重异构图注意力网络。载于 KDD 论文集。[27] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019。语言模型是无监督的多任务学习者。OpenAI 博客（2019）。[28] Filip Radlinski、Krisztian Balog、Bill Byrne 和 Karthik Krishnamoorthi。2019。指导式对话偏好需求引导：理解电影偏好的案例研究。载于 SIGDIAL 论文集。[29] Zhaochun Ren、Zhi Tian、Dongdong Li、Pengjie Ren、Liu Yang、Xin Xin、Huasheng Liang、Maarten de Rijke 和 Zhumin Chen。2022。关于对话式推荐中用户偏好的变分推理。载于 SIGIR 论文集。[30] Ananya B Sai、Akash Kumar Mohankumar、Siddhartha Arora 和 Mitesh M Khapra。2020。利用多参考对抗数据集和大规模预训练改进对话评估。《计算语言学协会学报》（2020）。[31] Fábio Souza、Rodrigo Nogueira 和 Roberto Lotufo。2019。使用 BERT-CRF 进行葡萄牙语命名实体识别。1–19。arXiv:1909.10649 http://arxiv.org/abs/1909.10649[32] Chongyang Tao、Lili Mou、Dongyan Zhao 和 Rui Yan。2018。Ruber：一种用于开放域对话系统自动评估的无监督方法。载于 AAAI 论文集。[33] Quan Tu、Shen Gao、Yanran Li、Jianwei Cui、Bin Wang 和 Rui Yan。2022。通过层次化信息建模实现对话式推荐。载于 SIGIR 论文集。[34] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017。Attention is All you Need。载于 NeurIPS 论文集。[35] Hu Xu、Seungwhan Moon、Honglei Liu、Bing Liu、Pararth Shah、Bing Liu 和 Philip Yu。2020。对话式推荐中的用户记忆推理。载于 COLING 论文集。[36] Yongfeng Zhang、Xu Chen、Qingyao Ai、Liu Yang 和 W. Bruce Croft。2018。迈向对话式搜索与推荐：系统提问，用户回应。载于国际信息与知识管理会议论文集。[37] Kun Zhou、Xiaolei Wang、Yuanhang Zhou、Chenzhan Shang、Yuan Cheng、Wayne Xin Zhao、Yaliang Li 和 Ji-Rong Wen。2021。CRSLab：用于构建对话式推荐系统的开源工具包。载于 ACL 论文集。[38] Kun Zhou、Wayne Xin Zhao、Shuqing Bian、Yuanhang Zhou、Ji-Rong Wen 和 Jingsong Yu。2020。通过基于知识图谱的语义融合改进对话推荐系统。载于 KDD 论文集。[39] Kun Zhou、Yuanhang Zhou、Wayne Xin Zhao、Xiaoke Wang 和 Ji-Rong Wen。2020。迈向主题引导的对话式推荐系统。载于 COLING 论文集。[40] Yuanhang Zhou、Kun Zhou、Wayne Xin Zhao、Cheng Wang、Peng Jiang 和 He Hu。2022。C^2-CRS：用于对话式推荐系统的由粗到细的对比学习。载于 WSDM 论文集。[41] Jie Zou、Evangelos Kanoulas、Pengjie Ren、Zhaochun Ren、Aixin Sun 和 Cheng Long。2022。通过基于 Transformer 的序列建模改进对话推荐系统。载于 SIGIR 论文集。